DOCUMENT RESUME

ED 167 608 TM 006 417

AUTHOR Hubert, John A.
TITLE A Review of the ESEA Title I Evaluation and Reporting System.
PUB DATE Mar 78
NOTE 34p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Achievement Gains; Achievement Tests; *Compensatory Education Programs; *Control Groups; Educationally Disadvantaged; Evaluation Methods; Federal Programs; Norm Referenced Tests; Post Testing; Pretesting; *Program Evaluation; *Research Design; *Research Problems; Scoring Formulas
IDENTIFIERS *Elementary Secondary Education Act Title I; Normal Curve Equivalent Scores; Regression Discontinuity Model

ABSTRACT

The Elementary and Secondary Education Act Title I Evaluation and Reporting System is a method for giving a federally funded project in reading or math an overall score on its cognitive effectiveness. This System introduced the Normal Curve Equivalent (NCE) as an aid in aggregating Title I program scores across states and nationwide regardless of specific tests used, or research design. NCE is used to measure Treatment Effect, defined as the difference between expected no treatment performance, and observed post-treatment performance of Title I participants. As a T-score the NCE is psychometrically sound and intelligible to lay people; more significantly, it indirectly eliminates grade equivalent scores. The System specifies three research designs each having a different control group: (1) classic pretest - posttest; (2) regression-discontinuity; and (3) norm-referenced. A basic criticism of all design is that they require the use of norm-referenced tests, but the validity of such tests has not been demonstrated for Title I students. Each of the designs has its own problems and merits. The Evaluation and Report System has achieved its goal of organizing and detailing evaluation procedures so that untrained evaluators can avoid most of the pitfalls; nevertheless, a highly trained evaluator is still needed, except in the simplest of programs. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

- ED167608

A REVIEW OF THE ESEA TITLE I EVALUATION & REPORTING SYSTEM

John A. Hubert
March 1978

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John A. Hubert

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM008 417

INTRODUCTION

Title I of the Elementary and Secondary Education Act funds programs in compensatory education. Under the regulations, these programs serve low-achieving students who are enrolled in Title I schools, i.e., schools with high concentrations of pupils from low-income families.

Since its inception in 1965 the Act has required local projects to do evaluations. There was considerable latitude for each project to conduct evaluations in its own way. There were several consequences from this. First, many evaluations were of poor technical quality, i.e., results were not reliable. Second, it was impossible to aggregate the results of local projects at the state and national level to determine net impact. Third, specially commissioned evaluations at the state and national level generally found modest levels of success, if any. The most famous of these, the Westinghouse evaluation of Project Head Start,[1] was debated nationally for both its methods and its findings, which were fairly dismal. Last, the Office of Education decided to develop a new evaluation system. That system, completely defined, is currently being piloted in 35 communities in Connecticut and elsewhere in the country. It is the subject of this discussion. The focus of discussion will be the methods for evaluating cognitive achievement, which is the central purpose of the funding.

The new system was carefully prepared. The RMC Research Corporation was contracted by the Office of Education to study the problem. In 1975 its results were published in two booklets[2] which described five research design models and various rules for selecting and implementing them. The Office of Education dropped one of the five (general regression) and consolidated two others in the process of finishing the System.

The Title I Evaluation and Reporting System, designed primarily for reading and math programs, includes requirements and procedures for systematic reporting on cognitive impact, pupil participation, Parent Advisory Councils, costs, personnel, and training. Guidelines for evaluation and reporting on these areas are provided in great detail in a manual and collection of booklets. Detailed instructions and forms are provided for making step by step calculations as well as for making judgments about the selection of tests, testing procedures, and choosing research design models.

No special instructions are provided for evaluation of affective or other goals, or for process evaluation. These are to be conducted according to the lights of the Local Education Agency.

Overall, the system includes the following major changes:

1) A system more tightly defined in its procedures and methods.
2) A complete set of handbooks, forms, etc., for the user to guide implementation.

3) A catalogue of three research designs.

4) A new unit of measure for assessing cognitive impact.

These last two items are the subject of discussion.

THE PURPOSE OF THE SYSTEM

There appears to be two basic goals which the evaluation system aims to accomplish. Both of these goals result from the establishment of a fairly specific and restricted approach to impact assessment.

1) With specified approaches and detailed guideline for implementation, the system will educate many persons cast into evaluation roles who are not especially well trained in the field. It will correspondingly tend to raise the level of evaluations being conducted in many districts, and make evaluation practice more uniform. Whether this has any drawbacks as well, is a question to be addressed.

2) The system provides a method for aggregating program scores across states and across the nation, regardless of the specific tests used or of the research design. This is done by measuring Treatment Effect with a common metric called the Normal Curve Equivalency or NCE. The system of scoring and aggregating scores will make it feasible to report to the nation on a program such as Head Start, either showing success or providing evidence to divert funds elsewhere.

I. Description of the System

Measurement, Scoring, and National Impact

In this System, the Office of Education offers a method for giving a national project in reading or math an overall score on its cognitive effectiveness. Obtaining a national aggregate score assumes that each local participant in the project follows a set of guidelines. The scores being aggregated need to be measurements of programs which are qualitatively alike.

The cognitive impact of the project is called the Treatment Effect.

The definition of Treatment effect is:

$$\left[\begin{array}{c}\text{Treatment}\\ \text{Effect}\end{array}\right] = \left[\begin{array}{c}\text{Observed}\\ \text{Post-treatment}\\ \text{Performance}\end{array}\right] - \left[\begin{array}{c}\text{Expected}\\ \text{No-treatment}\\ \text{Performance}\end{array}\right]$$

The observed post-treatment performance is always the mean or median posttest score of pupils participating in the project. The expected no-treatment performance is the post test mean score of a control group. One of three types of control group may be used, depending on the research design model.

Computation of Treatment Effect is as follows:

a) obtain mean standard score for each term on the right hand side of the equation above, according to the procedures dictated by the model used.

b) using the mean standard scores, obtain the two corresponding percentile scores.

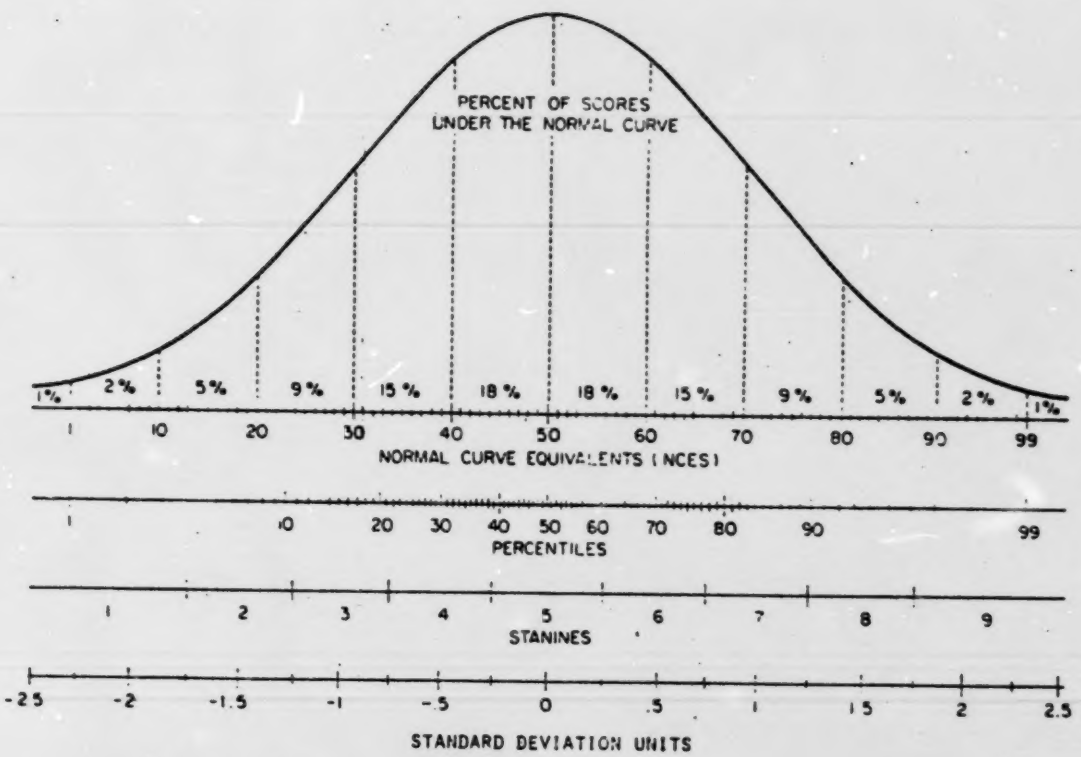

Figure 1. Area under the normal curve divided into NCEs, percentiles, stanines, and standard deviation units.

Note: Figure reproduced from the ESEA Title I Evaluation and Reporting System manual, p 9

c) for each percentile score, enter a table and obtain the NCE scores. Subtract to obtain the Treatment Effect. Positive scores are good, zero is no effect, and scores less than zero are bad.

The NCE is a t-score with a mean of 50.0 and a standard deviation of 21.06. The range is 1 to 99. It is thus an equal interval score which is equal to the percentile score at 1, 50, and 99. A universal table converts percentile to NCE.

The NCE is thus designed to build upon the public's experience with percentile scores. It is apparently intended that the NCE will eventually replace the grade equivalent score in the public and professional mind.

d) Having obtained an NCE gain for each school and grade, the local agency calculates a weighted NCE gain for the district.

$$NCEG_{dist.} = \frac{\Sigma NCEG \times N}{\Sigma N}$$

e) State and federal aggregate scores are also weighted averages of contributing scores.

With this background, we turn to describe the three research design models.

The Pretest - Posttest Control Group Design

This classic design emerged between 1900 and 1920 as a paradigm for research because it controlled for seven of the rival hypotheses listed by Campbell and Stanley.[3] Their design is used here to answer the question, "How much more did participants learn than they would have learned without the program?"

The question is answered with the difference in NCE score between the Title I mean posttest score and that of the control group.

If the pupils are randomly assigned to the two groups, the unadjusted difference in posttest mean scores may be used.

If small differences exist, the posttest scores may be adjusted using covariance, with the pretest scores as the covariate. This is done whether random assignment is used or not.

If the pre-test differences are larger than 5 NCE's and the pupils are not randomly assigned, then principle-axis adjustment is used. This is much the same as covariance but the adjustments are larger.

Basic rules of the model:

. Select a control group by the same procedures as the treatment group (e.g., randomly, bottom fifth of class). Be satisfied that controls are at least qualified for treatment.

- Test both groups at the same time on both pretest and posttest.
- If a non-normed test is used, a normed test must also be given to the treatment group some time during the year to get NCE scores. Thus a nationally normed instrument does not have to be the principal instrument, but one must be used at some point in the evaluation. This variation on the basic design is available for all three models.

Regression-discontinuity Model

Philosophically, the regression discontinuity model is attractive for Title I programs. Unlike the classic pretest - posttest control group model, it does not assume an equivalent control group. It assumes that the most educationally disadvantaged pupils are selected for the program and then compared in a special way with pupils who did better on a pre-test which was also used for selection of pupils.

Stanley and Campbell described this design for use with programs that select high-scoring pupils. In Title I, the design is used for programs that select low-scoring pupils.

The first step in the model is to conceptualize a population from whom participants will be selected, for instance, a school.

Second, a pretest is given to the entire population to determine entry into the program. All those who score below a cutoff score are in, and those above are not in. (If any exceptions to this rule are made, their scores must be deleted from the evaluation.)

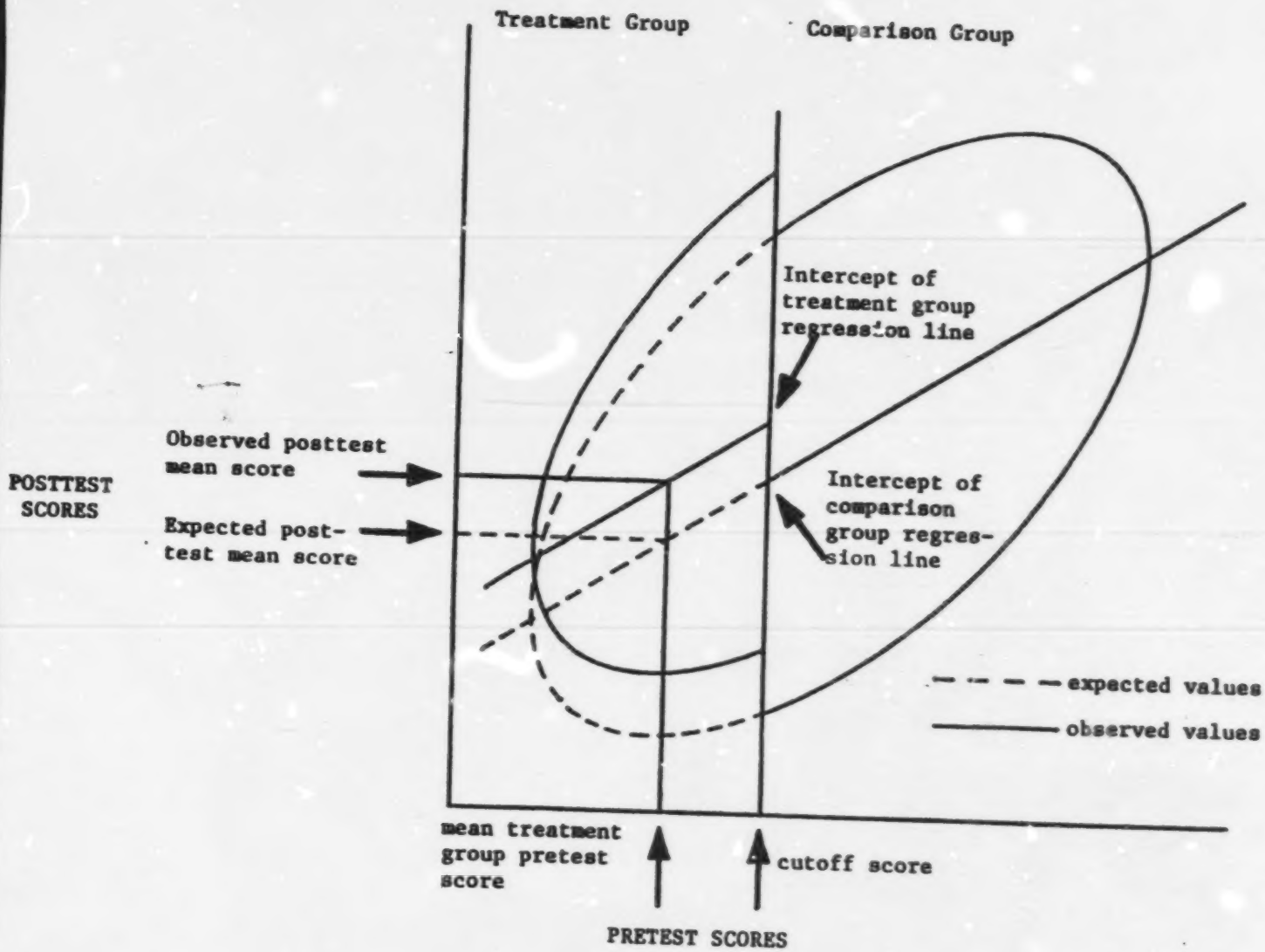

Fig. 2. Score distributions with treatment effect independent of pretest status. Figure reproced from "A Practical Guide...." 2

-10-

After the program is administered, a post test is given to both groups and comparison is made using a regression procedure.

This can best be understood by making a scatter plot of pre and post test scores. If there is no special program, then when regression lines are calculated for each group (above and below the cutoff score), the two lines tend to be co-terminus or the same line.

If the lower scoring pupils are in a special program that is effective, then their regression line will be higher. The difference in the height of the regression lines is a measure of program effect or Treatment Effect.

The design is unique because it requires that pupil selection be based upon the pretest scores alone. In most other designs, using the pre test for pupil selection is prohibited since it would evoke unwanted regression effects. Pupils scoring lower on the pre test will automatically tend to show more growth than pupils scoring higher on the pre test. That difference in growth will tend to poison the comparison between the two.

However, in the regression discontinuity design, growth due to regression is controlled for automatically.

The pre test does not have to be the same instrument as the post test, but should correlate highly with it. The pre test could be a composite measure designed for student selection.

Rules for implementing the regression discontinuity designs include:

- testing both groups at the same time, pre and post.
- using the pre test as the sole criterion for entry into the program.

Norm-referenced Design

The norm referenced design is not in the Stanley and Campbell list of experimental and quasi experimental designs. It is a design that assumes that pupils will tend to hold percentile rank unless their curriculum is changed. An increase in rank is evidence of the benefit of a new program.

When a nationally normed pre test is taken by program pupils, the design conceives norm pupils at the same percentile rank as constituting a control group. The control group is expected to maintain percentile rank at post testing, and that raw or standard scores become the expected no-treatment performance.

For instance, if program pupils score at the 19%ile in the fall, then their spring score will be compared to the standard score equivalent of 19%ile in the spring norms.

Testing must be conducted on or near the norm testing dates. As a rule of thumb, the Technical Assistance Center advises that the program testing must be within six weeks of the norm testing dates.

The user is also advised to be sure that the program population is like the norm population.

A big advantage of the model is that only program pupils need be tested. Moreover, computations are very simple.

One has but to determine 1) the pre test percentile and NCE scores; 2) the post test percentile and NCE scores; and 3) the difference in the NCE scores. A special t-test may also be conducted to determine whether the gain is significant.

Design Selection

The Evaluation System ranks the three designs and provides an algorithm for selection. In a nutshell:

1) If a random or equivalent control group is available, use Design B, the Pre test - Post test Control Group Design.
2) If Design C, regression discontinuity can be used, use it.
3) If not, use Design A, the Norm Referenced Design.

If none of the three designs are applicable, the system gives no advice, nor does it discuss other models. However, it does permit trained evaluators to use other sound models provided that it computes the Treatment Effect.

The system also encourages the user to supplement either Design B or C with Design A. The latter requires no additional testing if proper instruments and testing dates are selected.

II. Measurement Issues

The new unit of measure, the N.C.E., is just another T-score. From a purely mathematical viewpoint, it is perfectly sound, and has no impact upon what competent evaluators are already doing. It does, however, seem to carry psychological and strategical impact.

The N.C.E. builds upon the public's familiarity with percentiles, which is an interpretive advantage in making some approximate intuitive sense out of results, and that is an advantage. This possibility was the reason why this particular T-score was used. This similarity also raises the possibility of confusion between the two units. To the credit of the system, however, instructions are quite clear and complete, and this possibility does not appear to be a serious criticism.

More significantly, the introduction of the N.C.E. indirectly eliminates the grade equivalent score as an acceptable unit for reporting, and this result is most worthwhile. Hopefully, the change will be contagious.

The advantages of aggregating scores over projects are both political and technical, and this author has no objection to the aim. Under these circumstances, however, aggregation tends to conceal salient differences among the individual projects being aggregated. Even when the general plan and

objectives are shared, there are differences in instruments, pupils, research designs, curriculum, staff, and general climate under which the projects operate. It may be difficult for those concerned with politics to keep track of project-to-project differences in the face of tidy averages. The theoretical issue of averaging apples and oranges is closely tied with the practical problem of interpreting the answer.

The method of scoring makes no direct allowance for the quantity of service. The same treatment effect is calculated whether the project served students one hour a day or six. No distinction is made between projects operating several weeks or all year. However, the State does collect data on the amount and cost of services, and these will undoubtedly be factored into the analysis of aggregate results in some manner.

III. Instrumentation Issues

The norm-referenced design rightly requires that testing be conducted on or near the norming dates for the instrument. But few instruments have norming dates in both fall and spring. In fact, none of the instruments in the Anchor group have such norms; their fall-spring norms are interpolated. One new test on the market may have such norms. Similar problems are often found with more specialized tests. Thus persons who wish to use this model may not

be able to find any instruments which meet its requirements.

The system implicitly and explicitly encourages the use of standardized, norm-referenced tests, and yet the programs largely serve minority children. This combination raises the question of instrument validity. Validity is not a property of a test, but a relationship between a population, educational objectives, and test items and procedures. There is, therefore, no a priori reason to believethat the process of test development and validity established through and for white middle class children are equally valid for minority children. The fact of Title I programs is adequate testimony to the educational difference between these populations. The similarity between thedr ultimate educational goals should not convince us that the same items are necessarily capable of sampling the educational progress of children who approach those goals along different cultural paths. This principle is at the foundation of testing theory.

Another, related problem is the liklihood that many of the children in the norm sample at the level of program scores may also be in Title I or other similar programs. In that case, the assumed control group is actually another treatment group, or some unknown mix. Like the forgoing issue, this problem tends to be prejudicial to the Title I programs.

These criticisms are not new; Zimiles has noted that the sampling problems render the instruments useless because they do not measure the changes rendered. But they are particularly relevant to a critique of a system which was undertaken precisely because of a history of unexpected failure in these programs. It could well be that the many failures in Title I were due to faulty instruments, and not imprecision in the choice of statistical procedures. This possibility seems to be a fundamental oversight.

There is a need, in sum, for empirical investigation of the validity of norm-referenced instruments with respect to compensatory programs and their students and learning characteristics, when those instruments were developed with respect to mainstream children and programs.

IV. Research Design Issues

The procedures for aggregation of scores and other data provide information to state and federal officials to answer certain global questions. The most obvious of these are"How successful is the project as a whole?" and "Should the project as a whole be continued?". There are however, less obvious questions as well.

I recently surveyed evaluations of Head Start. One of my observations was that many evaluations of individual sites were favorable, whereas aggregate evaluation's tended to be mediocre. It appeared that aggregations were probably averaging strong and weak programs and ending with muddy results. The hazard is that many national projects could end up looking like failures because low-scoring sites offset the successes of high-scoring sites. It seems conceivable that we could get large funding cancellations if aggregate scores are not used intelligently.

A national assessment also needs to answer questions like, "What is the range of success of individual programs?" and "Which local programs should be discontinued or changed?" While this last question also has pitfalls, it does take advantage of having a common metric; it may stimulate local agencies to better performance, and, unlike a single aggregate score, it does not ignore salient differences among local programs.

The system has generally been designed to answer these questions too, as it systematically gathers information about costs, intensity of service, and the many other program variables beside project impact. But the interaction of politics with methodolgy is crucial, because widely differing pictures of results can be painted by including or excluding the variability between sites as a prominent aspect of aggregate findings.

We turn now to the individual research designs.

Stanley and Campbell are careful to list two designs separately: One is the classic pretest - posttest control group design with random assignment, and the other is the same design without random assignment. They make the distinction very carefully because non random assignment of pupils is full of pitfalls for the experimenter. Experimental and control pupils can perform differently for a wide variety of reasons having nothing to do with the program. Differences due to unequal regression may occur when the two groups belong to two distinct populations. When the two groups belong to two different schools, then history may affect them differently. For instance, disturbances may occur at one school which affect one group but not the other.

For continuing programs, as opposed to pilot programs, random assignment is not likely ever to be used in Title I. Since entry into the program is supposed to be based on need,

random assignment from equally needy pupils will not usually be consistent with program objectives.

For this reason it may have been wiser to aim the model straight at the likely implementation, anmely non-random assignment, and to focus instructions and guidelines on these applications. (Random assignment could have been a fourth model, but note comments following.) As it is, guidelines for the two cases are interwoven, and advice on non-random assignment is incomplete.

In advocating the use of comparable control groups in either of these cases, the system poses both a moral issue and a legal contradiction: If the pupils in the control group are comparable, they are also eligible, and therefore deprived of services to which the law and justice entitles them.

The regression-discontinuity design is a quasi-experimental model which has many strengths, and is particularly well-suited for use in Title I programs. However, a number of observations can be made regarding its use here.

The literature provides a number of variations on the Regression-discontinuity design. Stanley and Campbell point out the possibility of non-linear regression, and stress the importance of not assuming that the best fit is linear. Campbell also argues that

> "The pseudo effect produced by the slight degree of curvilinearity, and that so remote from the cutting point, warns against the general utility of these linear tests." 4

His research has apparently shown that small changes in the shape of the scatter-plot, even distant from the cut-off, affect the intersection of the regression line and the cut-off ordinate when the regression is linear. Thus the assessment of project impact may be erroneous.

Procedures are available for using random assignment at or near the cut-off point. It is also consistent with the basic concept to limit testing to just those pupils near the cut-off, provided that pre- and post- scores are obtained for an adquate number on both sides of it. In connection with this practice, it is well to remember the point made by Stanley and Campbell that the force of the regression discontinuity argument applies most strongly at the cut-off. As we move farther away, the possible rival hypotheses multiply. Among the other variation available are methods to analyze for discontinuity when the cut-off point is unknown or fuzzy.

The System addresses only the sharp cut-off method, despite the alternatives and their potential usefulness, and uses linear regression despite the warnings. A less obvious problem with regression discontinuity is a practical one: Since the model requires a particular method of pupil selection, it preempts a decision normally in the hands of others. Changing the pupil selection procedure will require the cooperation of parents, administration, and federal authorities. These obstacles will usually be sufficient to prevent the model from being used. This administration problem is a good reason

to open the door to more general regrassion methods. Any evaluator capable of doing regression-discontinuity is probably capable of other regression methods which don't have this cut-off requirement.

The System specifies that any instrument, rating, or composite measure which correlates highly with the posttest may be used as a pretest. By contrast, neither Campbell and Stanley, Boruch,[5] nor Iwanicki[6] mention correlation as a requirement. However, if there is no correlation, then there is no point to the method. The pre-score is no longer a true covariate, and the comparison is just betwen posttest means, as in a regular control group design.

The degree of correlation is not the matter of central concern. The method will be most successful when the pretest correlates highly with scholastic ability and previous relevant attainments, since its purpose is precisely to compensate for these systematic differences; and when the posttest is closely related to the treatment curriculum. If it isn't, the treatment group will not be able to show any improvement relative to a group superior in scholastic measures, regardless of correlation with a pretest.

The regression discontinuity model is a departure from the practice of using a median or mean score to represent the observed posttest performance. Instead, the model uses the projection of the regression line to the cut-off score. This practice is certainly in keeping with the nature of the model, but the intercept is no longer the mean or median or the group, but rather a higher score. This is slightly misleading in the context of reports on mean posttest scores, and the other models. Consistency could have been retained by using a technique suggested by Boruch.[7] He points out that the regression line of the control group can be projected across the cut-off to the mean prescore of the program pupils. This point would become the expected no-treatment performance. The method is not entirely sound, of course, but relies on the assumption already made, i.e., linearity of regression.

Since many educators do not favor exclusive reliance upon test scores for placement, an alternative is permissable: use of a rating of composite score. Stanley and Campbell

endorse this practice, as does Boruch. Iwanicki does not discuss composite eligibility scores. However, his discussion of the tie-breaking designs makes it evident that those designs are effective in avoiding one drawback of test scores, namely the error. Random assignment for those pupils within one standard error of the cut-off may be a simpler alternative than a composite measure, provided, of course, that test error is in fact the objection to the tests.

When a single measure is used, the System permits occasional students to be exempted from the cut-off rule. Such necessities sometimes occur. It requires in these instances that the scores of such students be completely deleted from the statistics. However, the System gives no guidelines to the permissable frequency of exceptions.

Another hazard is that some of the compensatory programs are designed to benefit both participants and non-participants. In some cases, the controls will be among the beneficiaries. For instance, those remaining in the regular classroom when program pupils are pulled out for supplementary services will enjoy a greater concentration of services from their teacher who, for the moment, has fewer pupils.

In the basic regression discontinuity design, selection and selection-related sources of invalidity are not usually a problem, but they can be. An example would be a small group of students with special problems or background, and extreme scores. These students are identified, and a program is funded which accomodates most or all of them. It could

thus turn out that, in such cases, students above the cut-off are different on a variety of dimensions, and that the total population is not homogeneous. In that case, selection, mortality, and selection interactions could invalidate the model. Iwanicki points out, however, that populations usually are homogeneous, and the above is not a problem.

A hazard overlooked by previous authors occurs when, as is usual, the pretest is a cognitive instrument. It is then possible for the student to sabotage the pretest in order to get into the program. As parents get acquainted with the program methods, this possibility seems real for programs that recur annually. Also, it is a fact of testing that low scores will include invalid or unresponsive results, regardless of motivation.

In summary, the System offers perhaps the simplest of several variations with which to implement regression discontinuity analysis. Worthwhile alternatives have been omitted, and, in assuming linearity, they have risked introducing unnecessary error. Finally, pupils may be selected for the program through invalid test scores.

The Norm Referenced Design is neither an experimental nor a quasi experimental design. It is a sliding criterion scale based upon testing experience with the norm, and upon experimental considerations. Its advantage lies in its convenience and ease of use, since no comparison group need be tested. The question is: How good is it? What are its flaws?

To the extent that certain conditions or assumptions are fulfilled, then the Norm Referenced Design more closely approximates a classical control group design and correspondingly more closely attains research validity. The principle conditions are as follows:

1) That the Title I population resembles the norm population at the levels of the pretest scores. That is, if the Title I population scores at the 20th percentile, that they are similar to 20th percentile pupils in the national sample.

But Title I pupils are mostly from low Socio-Economic-Status neighborhoods and have high concentrations of special ethnic or racial groups, as previously discussed.

Furthermore, times are different. Prevailing attitudes in the country as well as actual teaching practices have changed between the time the test was constructed, when it was normed, and when the model is used by a Title I program. For instance, this year's pupils who rank at the 25%ile among their contempories would rank lower, say 20%ile, on a test normed in 1963 when all students were scoring higher. This would have the effect of lowering the growth expectations for today's pupils.

2) The Norm-Referenced Model assumes that the norm sample population is not getting special treatment. Of course, Title I is such a large program, that a significant portion of pupils in the norm sample must also be recipients of Title I or similar programs if a recently normed test is used.

3) The Model assumes that trends like the contemporary declines in achievement scores across the nation do not affect the difference between Title I pre and post scores.

In short, the Model assumes that effects from the content of the test, sampling of the control group, regression, and history all cancel out, and that Title I pupils would hold percentile rank if they were'nt in the program.

4) In addition, it assumes that tests with suitable norms are available, as previously discussed.

5) It assumes that testing will be conducted very near to the norm testing dates. Selection of dates before or after the norming dates on either the pre test or post test will bias the results one way or the other. Discrepancies can be minimized by interpolating the norm tables, but it probably won't be done very often since the process may be tricky, depending on the format of the tables.

The aim of research design is to make sure that the experimental treatment is the only explanation for the results. The norm referenced design fails to control for any of the other explanation's but previous achievement, i.e., pretest score. Therefore it does not vaguely resemble an experimental design.

Against all these potential problems we may point to some advantages.

Previous to the norm referenced design, it was common to assume that a reasonable program goal was one month gain in score for each month in the program, measured in grade-equivalents.

We will not belabor here the many defects of the grade equivalent score as a unit of measure.[8,9] The important point is that this standard was heavily biased against programs which enrolled low-scoring students - precisely those whom Title I is most anxious to serve and help.

Second, the model is easy to administer and score. It does not require testing a control group, which saves time and money, and avoids problems arising from competition within the schools. It either frees an evaluator to do process evaluation or minimizes the need for a professional evaluator.

In summary, the model is defective as an experimental design. But it is an intuitively reasonable way to set achievement goals, it is simple to execute, it is an improvement over its predecessor, and there is some hope that the various defects will tend to cancel each other, and produce a reasonable result.

It is important to note that many programs exist which cannot readily use any of the three designs without substantial departures from the implementation requirements. These departures cast considerable doubt on the validity of the results. Perhaps only a minority of the programs fit comfortably into one of the three designs. At the same time, there are at least two other designs in Stanley and Capmbell that could well prove effective in many programs where none of the others fit. It is unclear why the System makes no provision for the use of less familiar designs such as time-series or the institutional cycle design.

In the course of my work with evaluation of Title I and similar programs in Hartford, it was my experience that each program was unique in some regard. A number of them presented special opportunities for little inventions or variations which have the effect of strengthening a design which was otherwise weak or defective. The System does not make any formal allowance for the use of such variations.

Thus the System presents us with the uncomfortable paradox of programs which call for highly individualized research designs being evaluated using a catalogue of only three. This paradox presumably arises because few districts are blessed with specialists in research design, and the evaluation won't, for the most part, use these other designs anyway.

There is a fundamental reason why it is difficult to find adequate experimental designs for these programs: They are not experimental programs. They are "ongoing" programs, and it is the basic intent of the funding legislation to serve the entire population in need. To the extent that the district succeeds, the possible control population is served by the program. This is one aspect of the situation which makes it difficult to establish experimental conditions.

Conclusions

1. The system has organized and detailed the evaluation procedures so well that untrained evaluators can surely avoid most of the common pitfalls. The detailed cookbooks are a noteworty accomplishment from an educational point of view.

2. The system has introduced a new unit of measure, the Normal Curve Equivalency, or NCE, which is psychometrically sound, intelligible to lay people, and which has the further advantage of displacing the grade equivalent score as an acceptable unit of measure for reporting results.

3. The first of the three research designs in the catalog is the familiar control group design, which is difficult to apply soundly when groups are not randomly assigned, as is almost always the case.

4. This same model contradicts the moral and legal aim of Title I by assuming comparable children not being served are available to serve as a control group.

5. The regression discontinuity design is potentially the most suitable, but requires a highly trained evaluator to implement. There are also political hazards with this model.

6. The norm referenced model is not a quasi-experimental model, but its advantages of simplicity and intelligibility make it potentially useful, but for instrumentation problems.

7. All models require the use of norm referenced instruments whose validity has not been demonstrated for Title I students and their objectives. It seems possible that this factor accounts for a major portion of the mediocre history of compensatory education.

8. The system overall has reduced the number of miscellaneous technical errors in evaluations, and therefore creates the impression of a scientific enterprise. But it has incorporated systematically the most likely cause of poor Title I results. The effect may be to lend more scientific credibility to the erroneous conclusion that Title I is worthless.

9. Despite its educational merits, a highly trained evaluator is probably still needed, except in the simplest of programs.

NOTES

(1) The study was criticized and defended in a group of articles reprinted from Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child. J.Hellmuth, Ed., New York: Brunner/Mazel, 1970
Authors: Campbell and Erlebacher; Victor G. Cicirelli; Evans and Schiller.

(2) A Practical Guide to Measuring Project Impact on Student Achievement 1975

A Procedure Guide for Validating Achievement Gains in Educational Projects 1976

Written by RMC Research Corporation,
Published by U.S. Office of Education

(3) Campbell, Donald T. and Stanley, Julian C. Experimental and Quasi-experimental Designs for Research on Teaching from the Handbook of Research on Teaching, Rand Mc Nally, 1963, pp 227-246.

(4) Campbell, Donald T. Measurement and Experimentation in Social Settings from a grant application.

(5) Boruch, Robert F. Three unpublished papers.

(6) Iwanicki,Edward F. A School Practitioner's Guide to Design for the Evaluation of Compensatory Educational Programs Research Report Series, Bureau of Educational Research and Service, School of Education, University of Connecticut.

(7) Boruch, Robert F. The Use of Linear Models in Evaluative Social Research Northwestern University 1973, p 11.

(8) Coleman, James S. and Korweit, Nancy L. Information Systems and Performances Measures in Schools, Educational Technology Publications, Englewood Cliffs, N.J., 1972.

(9) Wrightstone, J. Wayne; Hogan, Thomas P., Abbot, Muriel M., Accountability in Education and Associated Measurement Problems, in Test Service Notebook 33, Issued by the Test Department, Harcourt Brace Jovanovich, Inc.

(10) Zimiles, Herbert "A radical and regressive solution to the problem of evaluation" in Lillian and Katz, Current topics in Early Childhood Education, Ablex, 1977.